An Asymptotic Rate for the LASSO Loss

Cynthia Rush
Columbia University

Abstract

The LASSO is a well-studied method for use in high-dimensional linear regression where one wishes to recover a sparse vector $\boldsymbol{\beta} \in \mathbb{R}^p$ from noisy observations $\mathbf{y} \in \mathbb{R}^n$ measured through a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}$ where $\mathbf{w}$ is a vector of independent, mean-zero noise. We study the linear asymptotic regime where $n/p \to \delta$ for a constant $\delta \in (0, \infty)$. Using a carefully constructed approximate message passing (AMP) algorithm that converges to the LASSO estimator and recent finite sample theoretical performance guarantees for AMP, we provide large deviations bounds between various measures of LASSO loss and their concentrating values predicted by the AMP state evolution that shows exponentially fast convergence (in n) when the measurement matrix $\mathbf{X}$ is i.i.d. Gaussian. This work refines previous asymptotic analysis of LASSO loss in [Bayati and Montanari, 2012].

1 INTRODUCTION

A fundamental problem in high-dimensional statistics is estimating an unknown signal $\boldsymbol{\beta} \in \mathbb{R}^p$ from noisy measurements $\mathbf{y} \in \mathbb{R}^p$ in the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tag{1}$$

where $\mathbf{X}$ is a known $n \times p$ design matrix and $\mathbf{w} \in \mathbb{R}^p$ is vector of noise having independent entries and variance σ_z^2. Define $\delta = n/p > 0$. Strictly speaking, the 'high-dimensional' case is when $\delta \in (0, 1)$, meaning $n < p$, and this is the regime we study, although our results hold more generally when $\delta > 1$. Without any sort of structural or probabilistic constraints on the unknown signal $\boldsymbol{\beta}$, the high-dimensional problem is infeasible, so one often assumes that $\boldsymbol{\beta}$ is sparse, meaning most of its values are exactly equal to 0.

The LASSO [Tibshirani, 1996, Chen and Donoho, 1995], is a widely-used method for recovering an unknown, sparse signal $\boldsymbol{\beta}$ when given $\mathbf{X}$ and $\mathbf{y}$. It provides a sparse estimate computed as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b} \in \mathbb{R}^p} \mathcal{C}_\lambda(\mathbf{b}), \tag{2}$$

where, for a tuning parameter $\lambda > 0$,

$$\mathcal{C}_\lambda(\mathbf{b}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1, \tag{3}$$

where $\|\cdot\|_p$ denotes the standard ℓ_p-norm. Because of the penalty term $\lambda\|\mathbf{b}\|_1$, the LASSO estimate is sparse, with the level of sparsity controlled by the parameter λ. The LASSO reconstruction technique is well studied, and has been shown to provide good estimates of the truth $\boldsymbol{\beta}$, when $\boldsymbol{\beta}$ is sparse and the measurement matrix satisfies nice properties relating to the orthogonality of its columns on the support set of $\boldsymbol{\beta}$ (see, for example, [Hastie et al., 2015]).

This work studies asymptotically exact expressions for various types of loss, for example, the mean squared error (MSE), between the LASSO optimum $\widehat{\boldsymbol{\beta}}$ and the truth $\boldsymbol{\beta}$, where the asymptotics are with respect to the problem dimensions, what we refer to as the 'large system limit'[1]. Such characterizations of asymptotic loss for the LASSO have been studied both by developing Approximate Message Passing (AMP) algorithms that converge to the LASSO optimum [Bayati and Montanari, 2012] and by employing a convex Gaussian minmax theorem [Thrampoulidis et al., 2015]. In particular, in [Bayati and Montanari, 2012], it was shown that in the large system limit, the MSE, $p^{-1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$, is a deterministic value predicted by a fixed point of a scalar iteration in the case that the matrix $\mathbf{X}$ has i.i.d. Gaussian entries. The analysis builds from theoretical guarantees for the asymptotic performance of the efficient,

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

[1] The 'large system limit' refers to the setting where $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$

iterative AMP algorithms given in [Bayati and Montanari, 2011]. Recent results in [Rush and Venkataramanan, 2015] refine the asymptotic performance guarantees for AMP given in [Bayati and Montanari, 2011], by studying how the algorithm deviates from these precise asymptotic predictions in the non-asymptotic regime. However, for the non-asymptotic analysis, the AMP theory only extends to $O(\log n/ \log\log n)$ iterations, and the authors suggest that this rate may not be able to be improved with the current AMP theory that employs inductive proof methods.

We use the non-asymptotic analysis of AMP to study how the loss of the LASSO estimator deviates from its exact asymptotic performance predictions in the non-asymptotic regime, and show that the finite sample bound for the LASSO loss is of the same order as that of the finite sample bounds for AMP, thus extending to $O(\log n/ \log\log n)$ iterations. In particular, we provide a large deviations bound for the MSE, $p^{-1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ and other measures of loss for the LASSO estimator, in the case that the matrix $\mathbf{X}$ has i.i.d. Gaussian entries and the signal $\boldsymbol{\beta}$ has a prior distribution that its entries are i.i.d. according to some sub-Gaussian density p_B. This work refines the asymptotic results of [Bayati and Montanari, 2012] and shows that the rates of concentration for the LASSO loss match the rates at which the AMP iterates concentrate to predicted values provided by the AMP state evolution given in [Rush and Venkataramanan, 2015]. As in the previous asymptotic work, we prove this result by studying the efficient AMP algorithm, but now using the finite sample analysis of its performance.

Similar analyses of AMP algorithms have been used to provide asymptotic characterizations of the loss for a number of classical statistical estimation procedures including M-estimation [Donoho and Montanari, 2016], logistic regression [Sur and Candès, 2019], SLOPE [Bu et al., 2019], and ℓ_p-regularized least squares [Zheng et al., 2017], and also to study properties of these estimators, like in [Su et al., 2017]. We believe that the analysis pursued here can also be studied in these cases to give similar refinements of the asymptotic characterizations.

Notation. Bold, lower-case (upper-case) letters represent vectors (matrices). Non-bold letters denote scalars, as in v_i indexes the i^{th} element of the vector $\mathbf{v}$ and X_{ij} the $(i,j)^{th}$ element of $\mathbf{X}$. With a slight abuse of notation, upper-case letters represent random variables and lower-case letters their values, e.g. $Z \sim \mathcal{N}(\mu, \sigma^2)$ is a Gaussian random variable with mean μ and variance σ^2. We denote the set $\{1, 2, \dots, n\}$ by $[n]$. Throughout, $K, C, \kappa, c > 0$ are generic, positive constants whose values are not explicitly stated but do not depend on n or t. The indicator of event $\mathcal{A}$ is denoted $\mathbb{I}\{\mathcal{A}\}$.

Outline In Section 2, we introduce the AMP algorithms that converge to the LASSO optimum,and we also state the existing performance guarantees for such AMP algorithms. In Section 3 we present the main results, Theorem 2 and Theorem 3. Moreover, we prove Theorem 2 in Section 3 and then in Section 4 we state a technical result that is used to prove Theorem 3. We end with a discussion in Section 5.

2 AMP ALGORITHMS FOR LASSO

As mentioned previously, our analysis uses a class of approximate message passing (AMP) algorithms [Bayati and Montanari, 2011, Donoho et al., 2009a, Krzakala et al., 2012, Montanari, 2012, Rangan, 2011, Rangan et al., 2019] designed to converge to the LASSO optimum that were originally introduced in [Donoho et al., 2009b]. Before presenting the main results, in this section, we introduce the AMP algorithm in (5)-(6) and its theoretical performance guarantees provided by the state evolution given in (7).

AMP Definitions To iteratively update the estimate of the unknown vector, AMP uses the soft-threshold function, $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, defined as

$$\eta(x;\theta) = \text{sign}(x) \max\{|x| - \theta, 0\} \tag{4}$$

The soft-thresholding function equals 0 when the magnitude of its input is below some threshold, and shrinks the input towards 0 outside the threshold. Using the soft-thresholding function, the AMP algorithm updates as follows, iteratively providing estimates of the signal $\boldsymbol{\beta}^t \in \mathbb{R}^p$ and the residual $\mathbf{z}^t \in \mathbb{R}^p$. We initialize the AMP updates with $\boldsymbol{\beta}^0 = \mathbf{0} \in \mathbb{R}^p$ and for $t \geq 0$,

$$\boldsymbol{\beta}^{t+1} = \eta(\mathbf{X}^T\mathbf{z}^t + \boldsymbol{\beta}^t; \theta_t), \tag{5}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^t + \Big(\frac{\mathbf{z}^{t-1}}{\delta}\Big)\frac{\|\boldsymbol{\beta}^t\|_0}{p}, \tag{6}$$

where $\|\boldsymbol{\beta}^t\|_0 = \sum_{i=1}^p \mathbb{I}\{\beta_i^t = 0\}$ counts the number of non-zero values of $\boldsymbol{\beta}^t$, the soft-thresholding function acts element-wise on vector input, $\mathbf{X}^T$ denotes the transpose of the matrix $\mathbf{X}$, and $\{\theta_t\}_{t\geq 0}$ is a sequence of thresholds to be specified in what follows.

A remarkable property of AMP is that its performance can be characterized via a scalar recursion referred to as state evolution. Define a sequence $\{\tau_t^2\}_{t\geq 0}$ starting with $\tau_0^2 = \sigma^2 + \mathbb{E}\{B^2\}/\delta$ where $B \sim p_B$ (recalling that p_B specifies the prior distribution on the elements of the signal) and σ^2 is the noise variance, i.e. $\mathbb{E}[w_i^2] = \sigma^2$ for $i = 1, 2, \dots, n$. Then for $t \geq 0$, let

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big\{\Big[\eta(B + \tau_t Z; \theta_t) - B\Big]^2\Big\}, \tag{7}$$

where $Z \sim \mathcal{N}(0,1)$ independent of $B \sim p_B$ and $\delta = n/p$. AMP has the desirable property that the input to the soft-thresholding function, $\mathbf{X}^T\mathbf{z}^t + \boldsymbol{\beta}^t$ (see (5)), which we henceforth refer to as the 'effective observation' is approximately equal in distribution to the true signal plus independent Gaussian noise, where the variance of the noise is given by the state evolution. Namely, $\mathbf{X}^T\mathbf{z}^t + \boldsymbol{\beta}^t \approx \boldsymbol{\beta} + \tau_t \mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_{p \times p})$ where $\mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix. This approximation is made precise asymptotically in [Bayati and Montanari, 2011], and [Rush and Venkataramanan, 2015] shows that the deviation between the two, $\boldsymbol{\Delta}^t := (\mathbf{X}^T\mathbf{z}^t + \boldsymbol{\beta}^t) - (\boldsymbol{\beta} + \tau_t\mathbf{Z})$, concentrates to zero exponentially fast, in the sense that for $\epsilon \in (0,1)$,

$$P\Big(\frac{1}{p}\|\boldsymbol{\Delta}^t\|^2 \geq \epsilon\Big) \leq K_t e^{-\kappa_t p \epsilon^2},$$

for constants $K_t, \kappa_t > 0$ not depending on p or ϵ, but depending on t.

2.1 AMP Performance

We first present the performance guarantees for AMP from [Rush and Venkataramanan, 2015] that are used for our work. Their analysis uses a stopping criterion, set by the user, that terminates the algorithm once the expected squared error of the estimates is either very small or stops improving appreciably. The specific form of the stopping criterion is not important for our presentation, so we do not elaborate on its definition, however we note that the algorithm is run only for iterations $0 \leq t < T^*$ where the exact value of the stopping iteration T^* is discussed in the cited work.

The AMP performance analysis provided by Theorem 1, restated below from [Rush and Venkataramanan, 2015, Thm. 3.1], shows that the loss of the AMP estimates $\boldsymbol{\beta}^t$ at any iteration $0 < t < T^*$, when the loss is measured by a pseudo-Lipschitz loss function, concentrates to a deterministic value predicted by the state evolution for large, but finite n. A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz (of order 2) if there exists a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq L(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\|\mathbf{x} - \mathbf{y}\|$.

Finally, we make the following assumptions, which are needed for the AMP concentration result to hold.

(A1) Measurement Matrix: The entries of the $n \times p$ measurement matrix $\mathbf{X}$ are i.i.d. $\sim \mathcal{N}(0, 1/n)$.

(A2) Signal: The prior distribution of the length$-p$ signal $\boldsymbol{\beta}$ assumes the entries are i.i.d. according to a sub-Gaussian distribution p_B with $\mathbb{E}\{\beta_i^2\} = \sigma_0^2$ for $i \in [p]$ where we assume $0 < \sigma_0^2 < \infty$. The sub-Gaussian assumption implies [Boucheron et al., 2013], for generic constants $K, \kappa \geq 0$,

$$P\Big(\Big|\frac{1}{p}\|\boldsymbol{\beta}\|^2 - \sigma_0^2\Big| \geq \epsilon\Big) \leq K e^{-\kappa n \epsilon^2}. \tag{8}$$

(A3) Measurement Noise: The entries of the length$-n$ measurement noise vector $\mathbf{w}$ are i.i.d. according to some sub-Gaussian distribution p_W with mean 0 and $\mathbb{E}[w_i^2] = \sigma^2 < \infty$ for $i \in [n]$.

Below, we restate the AMP performance guarantees [Rush and Venkataramanan, 2015, Thm. 3.1] that are used in our analysis.

Theorem 1. *[Rush and Venkataramanan, 2015, Thm. 3.1] Under assumptions **(A1)** – **(A3)**, the following holds for any (order-2) pseudo-Lipschitz function $\phi : \mathbb{R}^2 \to \mathbb{R}$, $\epsilon \in (0,1)$, and $0 \leq t < T^*$, where T^* is determined by the stopping criterion.*

$$P\Big(\Big|\frac{1}{p}\sum_{i=1}^{p} \phi(\beta_i^t, \beta_i) - \mathbb{E}\Big[\phi\Big(\eta(B + \tau_t Z; \theta_t), B\Big)\Big]\Big| \geq \epsilon\Big) \\ \leq K_t e^{-\kappa_t n \epsilon^2}.$$

In the expectation above, $B \sim p_B$ and $Z \sim \mathcal{N}(0,1)$ are independent, and τ_t is given by (7). The constants K_t, κ_t are given by $K_t = C^{2t}(t!)^{10}, \kappa_t = \frac{1}{c^{2t}(t!)^{22}}$, where $C, c > 0$ are universal constants (not depending on t, n, or ϵ) that are not explicitly specified.

For the AMP algorithm introduced in (5)-(6) to converge to the LASSO optimum, one needs to carefully choose the sequence of thresholds $\{\theta_t\}_{t \geq 0}$. As introduced in [Donoho et al., 2009b], these thresholds ultimately control the sparsity of the AMP estimate in a similar manner to λ in the LASSO cost (3).

2.2 AMP Relation to the LASSO Solution

We connect the Theorem 1 performance guarantees to the LASSO solution (2), by relating the thresholds $\{\theta_t\}_{t\geq 0}$ to the LASSO parameter $\lambda > 0$ in the following way. Throughout, we take $\theta_t = \alpha \tau_t$ where $\alpha > 0$ is fixed and τ_t is given by the state evolution in (7). Now we restate the state evolution given in (7) with this choice of θ_t, noting that moving forward this is the state evolution that we will reference. Taking $\tau_0^2 = \sigma^2 + \mathbb{E}\{B^2\}/\delta$ where $B \sim p_B$ and σ^2 is the variance in (1), for $t \geq 0$ define $\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \alpha\tau_t)$ where

$$\mathsf{F}(\tau^2, \alpha\tau) = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big\{\Big[\eta(B + \tau Z; \alpha\tau) - B\Big]^2\Big\}, \tag{9}$$

where $Z \sim \mathcal{N}(0,1)$ independent of $B \sim p_B$. This is an intuitive choice of threshold: since the effective observation is the input to the soft-thresholding function in (5) and it is approximately distributed as $\boldsymbol{\beta} + \tau_t\mathbf{Z}$, it is natural to make the threshold proportional to the standard deviation of the noise, τ_t, in order to separate noise from signal. This choice of threshold has a number of nice properties that have been investigated in [Donoho et al., 2009b], among other works.

The recursion in (9) has been well-studied [Bayati and Montanari, 2012, Donoho et al., 2009b, Donoho et al., 2009a], and, as stated in the following lemma, it has been shown that the sequence $\{\tau_t^2\}_{t\geq 0}$ always converges to the solution of the fixed point equation

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big\{\Big[\eta(B + \tau Z; \alpha\tau) - B\Big]^2\Big\}, \tag{10}$$

for appropriately chosen α.

Lemma 1. *[Bayati and Montanari, 2012, Prop. 1.3] Let α_{min} to be the unique, nonnegative solution to $(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \delta/2$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the Gaussian CDF and PDF, respectively. Then for $\sigma^2 > 0$ and $\alpha > \alpha_{min}$, the state evolution mapping $\mathsf{F}(\tau^2, \alpha\tau)$ defined in (9) is concave and monotone increasing in τ^2. Moreover, there exists a unique τ_*^2 such that $\mathsf{F}(\tau_*^2, \alpha\tau_*) = \tau_*^2$ and the monotone sequence $\{\tau_t^2\}_{t\geq 0}$ converges to τ_*^2 as $t \to \infty$.*

The choice of α value in the AMP equations controls the sparsity of the solutions, serving an analogous role to λ in the LASSO cost. Therefore, to relate the AMP algorithm introduced in (5)-(6) to the optimum of the LASSO cost (3), we use a calibration between λ and α values provided in [Bayati and Montanari, 2012]. In particular, there is a one-to-one correspondence between λ in (3) and the choice of $\alpha_{min} < \alpha < \infty$ used in the threshold for the AMP updates in (5). The following relationship is used to determine the threshold level α used in the AMP iterates (5) - (6) for a given value $\lambda > 0$ in (3):

$$\lambda(\alpha) = \alpha\tau_*\Big(1 - \frac{1}{\delta}\mathbb{E}\Big[\eta'(B + \tau_* Z; \alpha\tau_*)\Big]\Big), \tag{11}$$

where $\eta'(\cdot)$ is the weak derivative of the soft-thresholding function $\eta(\cdot,\cdot)$ defined in (4), $Z \sim \mathcal{N}(0,1)$ is independent of $B \sim p_B$, and τ_* is defined in Lemma 1. We will need to invert the above function and we call the inversion $\alpha(\lambda)$. Details about $\alpha(\lambda)$ are given in [Bayati and Montanari, 2012], for this manuscript we simply need to know that such a function is well defined.

3 MAIN RESULT

The main result shows concentration for pseudo-Lipschitz loss functions of the minimizer of the LASSO cost function to the truth $\boldsymbol{\beta}$. This is a finite sample version of [Bayati and Montanari, 2012, Thm. 1.5]. Recall, τ_* denotes the unique fixed point solving (10) when α is selected according to the function $\alpha(\lambda)$ for a given value of λ. We know that the state evolution converges to τ_* as $t \to \infty$. We denote $\theta_* = \alpha\tau_*$.

The two main results, Theorem 2 and Theorem 3, are concentration results involving specific t-dependent (universal) constants, δ_t, ν_t. At the end of this section we state and prove Lemma 2, which defines these values explicitly and shows that both approach 0 as $t \to \infty$.

Theorem 2. *Assume, $t = O\Big(\frac{\log n}{\log\log n}\Big)$. For any (order-2) pseudo-Lipschitz function $\phi : \mathbb{R}^2 \to \mathbb{R}$, under the same assumptions as those in Theorem 1, for the LASSO solution in (2) and $t_{min} \leq t \leq T^*$ (for some $t_{min} < T^*$ specified in more detail in Section 4), define an event $\mathcal{T}(p)$ as*

$$\begin{aligned}\Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\widehat{\beta}_i, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)]\Big| \\ \geq \epsilon + \delta_t + \widetilde{\kappa}\nu_t,\end{aligned} \tag{12}$$

where $\widetilde{\kappa} > 0$ is a universal constant and ν_t and δ_t are defined in Lemma 2. Then, for large enough n,

$$P(\mathcal{T}(p)) \leq K K_t e^{-\kappa\kappa_t n\epsilon^2}. \tag{13}$$

Constants K_t, κ_t are defined in the Thm. 1 statement.

Remark 1. *To get a more concrete idea of the result in Theorem 2, it is useful to consider some specific pseudo-Lipschitz functions. First, let $\phi(a,b) = (a-b)^2$, then Theorem 2 proves that the MSE of the LASSO minimizer concentrates to a known constant value with exponentially small probability of error in n. In particular, for all $t \geq 0$,*

$$\begin{aligned}P\Big(\Big|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p - \delta(\tau_* - \sigma^2)\Big| \geq \epsilon + \delta_t + \widetilde{\kappa}\nu_t\Big) \\ \leq K K_t e^{-\kappa\kappa_t n\epsilon^2}.\end{aligned}$$

Similarly, taking $\phi(a,b) = |a-b|$ the theorem proves that the normalized L_1-error, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1/p$, concentrates around $\mathbb{E}|\eta(X + \tau_* Z; \theta_*) - X|$.

Remark 2. *The bound in Theorem 2 implies the asymptotic result of [Bayati and Montanari, 2012, Thm. 1.5]. To see this, notice that the sum $\sum_{p=1}^{\infty} P(\mathcal{T}(p)) \leq \sum_{p=1}^{\infty} K K_t e^{-\kappa\kappa_t\delta p\epsilon^2}$ is finite for all $t \geq 0$, and therefore*

$$\begin{aligned}\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\phi(\widehat{\beta}_i, \beta_i) \\ \overset{a.s.}{=} \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)] + \delta_t + \widetilde{\kappa}\nu_t.\end{aligned}$$

Taking the limit $t \to \infty$, the terms $\delta_t, \nu_t \to 0$ as shown in Lemma (2). We note that taking the limits in the *other direction, namely $t \to \infty$ followed by $p \to \infty$ would result in term that is greater than 1 on the right hand side of the upper bound provided in* (13).

Remark 3. *Theorem 2 also refines the asymptotic convergence result [Bayati and Montanari, 2012, Thm. 1.5] in that it specifies how large t can be (compared to the dimension n) for the state evolution predictions to be meaningful in characterizing AMP performance. As detailed in [Rush and Venkataramanan, 2015], using the expression for κ_t in the Theorem 1 statement, we need that $t = O\Big(\frac{\log n}{\log\log n}\Big)$ in order for the term in the exponent of* (13), *$\kappa\kappa_t n\epsilon^2 \to \infty$ as n grows.*

Thus, when the AMP is run for a growing number of iterations, the state evolution predictions are guaranteed to be valid until iteration t if the problem dimension grows faster than exponentially in t.

Theorem 2 is a direct consequence of the following result showing that the mean square error between the AMP estimate at any time t and the LASSO solution concentrates on some deterministic value that is decreasing to 0 with t. This is a large deviations version of [Bayati and Montanari, 2012, Thm. 1.8].

Theorem 3. *Assume that $t = O\Big(\frac{\log n}{\log\log n}\Big)$. For the LASSO solution in* (3) *and $\{\boldsymbol{\beta}^t\}_{0\leq t\leq T^*}$, the estimate of the AMP iteration at time $t_{min} \leq t \leq T^*$ (for some $t_{min} < T^*$ specified in more detail in Section 4), we have for $\epsilon \in (0,1)$, and large enough n,*

$$P\Big(\frac{1}{p}\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\|^2 \geq \epsilon + \nu_t^2\Big) \leq KK_t e^{-\kappa\kappa_t n\epsilon^2}, \qquad (14)$$

where ν_t^2 is defined in Lemma 2. The constants K_t, κ_t are given by $K_t = C^{2t}(t!)^{C_1}, \kappa_t = \frac{1}{c^{2t}(t!)^{c_1}}$, where $C, C_1, c, c_1 > 0$ are universal constants (not depending on t, n, or ϵ) that are not explicitly specified.

We next use Theorem 3 to prove Theorem 2 and then we prove Theorem 3 in what follows.

Proof of Theorem 2. Let $\widetilde{\kappa} = 2L\sqrt{5 + 4\sigma_0^2 + 2\mathsf{b}_1}$ where σ_0^2 is the element-wise variance of the signal defined in (8), L is the pseudo-Lipschitz constant of ϕ, and $\mathsf{b}_1 = \max\{\hat{\mathsf{B}}, \tilde{\mathsf{B}}\}$ are values defined in Lemma 7 in the supplementary materials that are high-probability upperbounds on the standardized norms of the AMP estimate $\boldsymbol{\beta}^t$ and the LASSO minimizer $\widehat{\boldsymbol{\beta}}$. Now, considering the event $\mathcal{T}(p)$ defined in (12), we see that

$$\begin{aligned}&\Big|\frac{1}{p}\sum_{i=1}^{p}\phi([\widehat{\beta}_i, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)]\Big| \\ &\leq \frac{1}{p}\sum_{i=1}^{p}\Big|\phi(\widehat{\beta}_i, \beta_i) - \phi(\beta_i^t, \beta_i)\Big| \\ &\qquad + \Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\beta_i^t, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)]\Big|.\end{aligned}$$

Therefore, $P(\mathcal{T}(p))$ is upper bounded by

$$\begin{aligned}&P\Big(\Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\beta_i^t, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)]\Big| \geq \frac{\epsilon}{2} + \delta_t\Big) \\ &+ P\Big(\frac{1}{p}\sum_{i=1}^{p}\Big|\phi(\widehat{\beta}_i, \beta_i) - \phi(\beta_i^t, \beta_i)\Big| \geq \frac{\epsilon}{2} + \widetilde{\kappa}\nu_t\Big). \qquad (15)\end{aligned}$$

Label the two terms of (15) as T_1 and T_2. We provide upper bounds for both.

First consider T_1 of (15). Recalling the definition of δ_t from (18), notice that

$$\begin{aligned}&\Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\beta_i^t, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_* Z; \theta_*), B)]\Big| \\ &\leq \Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\beta_i^t, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_t Z; \theta_t), B)]\Big| + \delta_t\end{aligned}$$

Therefore, by Theorem 1, we can upper bound T_1 with

$$\begin{aligned}&P\Big(\Big|\frac{1}{p}\sum_{i=1}^{p}\phi(\beta_i^t, \beta_i) - \mathbb{E}[\phi(\eta(B + \tau_t Z; \theta_t), B)]\Big| \geq \frac{\epsilon}{2}\Big) \\ &\qquad \leq K_t e^{-\kappa_t n\epsilon^2/4}.\end{aligned}$$

Now consider term T_2 of (15). First notice that by the Triangle Inequality and the the pseudo-Lipschitz property of ϕ, for all $i \in [p]$, we have

$$|\phi(\widehat{\beta}_i, \beta_i) - \phi(\beta_i^t, \beta_i)| \leq L(1 + 2|\beta_i| + |\widehat{\beta}_i| + |\beta_i^t|)|\widehat{\beta}_i - \beta_i^t|.$$

Then by Cauchy-Schwarz,

$$\begin{aligned}&\Big(\sum_{i=1}^{p}\Big|\phi(\widehat{\beta}_i, \beta_i) - \phi(\beta_i^t, \beta_i)\Big|\Big)^2 \\ &\leq L^2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2\sum_{i=1}^{p}(1 + 2|\beta_i| + |\widehat{\beta}_i| + |\beta_i^t|)^2.\end{aligned}$$

Finally we notice that by Cauchy-Schwarz again,

$$\sum_{i=1}^{p}(1+2|\beta_i|+|\widehat{\beta}_i|+|\beta_i^t|)^2 \leq 4(p+4\|\boldsymbol{\beta}\|^2+\|\widehat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\beta}^t\|^2).$$

Therefore, term T_2 is upper bounded by

$$P\Big(\frac{2L\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|}{\sqrt{p}}\sqrt{1+\frac{4\|\boldsymbol{\beta}\|^2+\|\widehat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\beta}^t\|^2}{p}}\geq\frac{\epsilon}{2}+\widetilde{\kappa}\nu_t\Big). \tag{16}$$

Notice that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|/\sqrt{p}$ concentrates to ν_t by Theorem 3 and the terms under the square root concentrate to constants by (8) and Lemma 7 in the supplementary materials. We now show how this leads to an upper bound on the probability in (16).

Recall, $\widetilde{\kappa}=2L\sqrt{5+4\sigma_0^2+2\mathsf{b}_1}$. Then if the following event is true,

$$\Big\{\frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|}{\sqrt{p}}\leq\nu_t+\frac{\epsilon}{2\widetilde{\kappa}}\Big\}\cap\Big\{\frac{\|\boldsymbol{\beta}\|^2}{p}\leq\sigma_0^2+\epsilon\Big\}$$
$$\cap\Big\{\frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p}\leq\mathsf{b}_1\Big\}\cap\Big\{\frac{\|\boldsymbol{\beta}^t\|^2}{p}\leq\mathsf{b}_1\Big\},$$

it follows that

$$\begin{aligned}&\frac{2L\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|}{\sqrt{p}}\sqrt{1+4\frac{\|\boldsymbol{\beta}\|^2}{p}+\frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p}+\frac{\|\boldsymbol{\beta}^t\|^2}{p}}\\&\leq 2L\Big(\nu_t+\frac{\epsilon}{2\widetilde{\kappa}}\Big)\sqrt{1+4(\sigma_0^2+\epsilon)+2\mathsf{b}_1}\\&\leq\Big(\nu_t+\frac{\epsilon}{2\widetilde{\kappa}}\Big)\widetilde{\kappa}\leq\frac{\epsilon}{2}+\widetilde{\kappa}\nu_t.\end{aligned}$$

Now we use this to bound the probability in (16). Notice, for events A, B, and C, if $A\cap B\implies C$ then $P(C)\geq P(A\cap B)$ and $P(C^c)\leq P((A\cap B)^c)=P(A^c\cup B^c)\leq P(A^c)+P(B^c)$. Therefore,

$$\begin{aligned}&P\Big(\frac{2L\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|}{\sqrt{p}}\sqrt{1+4\frac{\|\boldsymbol{\beta}\|^2}{p}+\frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p}+\frac{\|\boldsymbol{\beta}^t\|^2}{p}}\geq\epsilon+\widetilde{\kappa}\nu_t\Big)\\&\leq P\Big(\frac{\|\boldsymbol{\beta}\|^2}{p}\geq\sigma_0^2+\epsilon\Big)+P\Big(\frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p}\geq\mathsf{b}_1\Big)\\&\quad+P\Big(\frac{\|\boldsymbol{\beta}^t\|^2}{p}\geq\mathsf{b}_1\Big)+P\Big(\frac{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^t\|}{\sqrt{p}}\geq\nu_t+\frac{\epsilon}{2\widetilde{\kappa}}\Big)\\&\leq Ke^{-\kappa n\epsilon^2}+Ke^{-\kappa n\epsilon^2}+Ke^{-\kappa n\epsilon^2}+KK_te^{\frac{-\kappa\kappa_t\nu_t^2n\epsilon^2}{4}}.\end{aligned}\tag{17}$$

The final inequality follows from (8), Lemma 7 in the supplementary materials, and Theorem 3 along with Lemma 1 in the supplementary materials. We note that the use of Lemma 1 in the supplementary materials and Theorem 3 means thats there is a ν_t^2 in the concentration bound of (17) and therefore in (13). However, noting that κ_t has a $\frac{1}{t!}$ term, the fact that ν_t decays like $te^{-\kappa t}$ as shown in Lemma 2 means that the presence of ν_t^2 in the bound won't change the overall rate of the concentration (meaning how t can grow with n), so it is not explicitly stated. Similarly, for the additional t in front. □

Finally, we explicitly define the t-dependent constants that show up in the two theorems above and we show their rates of convergence.

Lemma 2. *Define the following t-dependent terms. For any (order-2) pseudo-Lipschitz function $\phi:\mathbb{R}^2\to\mathbb{R}$, define for $t\geq 0$,*

$$\delta_t:=\tag{18}$$
$$\Big|\mathbb{E}\Big[\phi(\eta(B+\tau_*Z;\theta_*),B)-\phi(\eta(B+\tau_tZ;\theta_t),B)\Big]\Big|,$$
$$\widehat{e}_{t+1}:=\Big|\lambda-\theta_t\Big[1-\frac{1}{\delta}\mathbb{E}[\eta'(B+\tau_tZ;\theta_t)]\Big]\Big|,\tag{19}$$

where $Z\sim\mathcal{N}(0,1)$ independent of $B\sim p_B$ and for constant $\kappa_1>0$,

$$\begin{aligned}e_t:=\Big[\frac{\lambda^2}{\theta_{t-1}^2}+c_{max}\Big]\delta(\tau_t^2-2E_{t,t-1}+\tau_{t-1}^2)\\+\frac{(\widehat{e}_t)^2\delta c_{max}}{\alpha^2}\end{aligned}\tag{20}$$

$$\nu_t^2:=\kappa_1e_t\Big[\frac{(1+tc_{max})e_t}{\lambda^2}+\frac{2+tc_{max}}{c_{min}}\Big],\tag{21}$$

where c_{min} and c_{max} are the concentrating values for the maximum and minimum (non-zero) singular values of the matrix $\mathbf{X}$ as defined in Lemma 4, Condition (4), and $E_{t+1,t}$ is defined by the following more general state evolution recursion. Define a set of covariances, $\{E_{r,s}\}_{r\geq0,s\geq0}$ recursively such that

$$\begin{aligned}&\delta E_{s+1,r+1}=\delta\sigma^2+\\&\mathbb{E}\left\{[\eta(B+\tau_sZ_s;\theta_s)-B][\eta(B+\tau_rZ_r;\theta_r)-B]\right\},\end{aligned}\tag{22}$$

where Z_s and Z_r are jointly Gaussian but independent of $B\sim p_{\boldsymbol{\beta}}$, with $\mathbb{E}[Z_s]=\mathbb{E}[Z_r]=0$, $\mathbb{E}[Z_s^2]=\mathbb{E}[Z_r^2]=1$, and $\mathbb{E}[Z_sZ_r]=\frac{E_{s,r}}{\tau_r\tau_s}$. Note that $E_{t,t}=\tau_t^2$ defined in (7). Finally we define the boundaries such that $E_{0,0}=\sigma^2+\frac{1}{\delta}\mathbb{E}\left\{B^2\right\}=\tau_0^2$ and

$$\delta E_{0,t+1}=\delta\sigma^2+\mathbb{E}\left\{[-B][\eta(B+\tau_tZ_t;\theta_t)-B]\right\},\tag{23}$$

for $Z_t\sim\mathcal{N}(0,\tau_t^2)$ independent of $B\sim p_{\boldsymbol{\beta}}$.

Then δ_t, e_t, and $\widehat{e}_t$ converge to 0 (as t grows) like $Ke^{-\kappa t}$ while ν_t^2 converges to 0 like $Kte^{-\kappa t}$.

Proof. The proof uses the exponential convergence of the state evolution sequence proved in [Bayati and Montanari, 2012, Appendix C] and restated below in Lemma 3 for convenience. The proof of Lemma 3 doesn't change for our analysis and is quite technical, so we don't include the details.

Lemma 3. *[Bayati and Montanari, 2012, Lemma 5.7] Assume $\alpha>\alpha_{min}$ defined in Lemma 1 and let $\{E_{s,t}\}_{s,t\geq0}$ be defined by recursion (22) with initial condition (23). Then, for all $t\geq0$, there exist constants $B_1, r_1>0$ such that, $|E_{t,t}-\tau_*^2|\leq B_1e^{-r_1t}$,* *and* $|E_{t,t+1} - \tau_*^2| \leq B_1 e^{-r_1 t}$. *Moreover,* $E_{t-1,t-1}^2 - 2E_{t-1,t} + E_{t,t}^2 \leq B_1 e^{-r_1 t}$.

We first prove, for universal constants $K, \kappa > 0$, that $\delta_t \leq K e^{-\kappa t}$. To see this, recall that $\phi(\cdot,\cdot)$ is pseudo-Lipschitz, and therefore,

$$\begin{aligned}
&\Big|\phi(\eta(B + \tau_* Z; \theta_*), B) - \phi(\eta(B + \tau_t Z; \theta_t), B)\Big| \\
&\leq L\Big[1 + 2|B| + |\eta(B + \tau_* Z; \theta_*)| + |\eta(B + \tau_t Z; \theta_t)|\Big] \\
&\qquad \times \Big|\eta(B + \tau_* Z; \theta_*) - \eta(B + \tau_t Z; \theta_t)\Big|. \quad (24)
\end{aligned}$$

For the soft-thresholding function defined in (4), $|\eta(x,\theta)| \leq |x|$ and it is easy to show the following bounds $|\eta(x_1,\theta) - \eta(x_2,\theta)| \leq |x_1 - x_2|$ and $|\eta(x,\theta_1) - \eta(x,\theta_2)| \leq |\theta_1 - \theta_2|$. This implies

$$\begin{aligned}
&\Big|\eta(B + \tau_* Z; \theta_*) - \eta(B + \tau_t Z; \theta_t)\Big| \\
&\leq \Big|\eta(B + \tau_* Z; \theta_*) - \eta(B + \tau_* Z; \theta_t)\Big| \\
&\quad + \Big|\eta(B + \tau_* Z; \theta_t) - \eta(B + \tau_t Z; \theta_t)\Big| \\
&\leq |\theta_* - \theta_t| + |\tau_* - \tau_t||Z| = |\tau_* - \tau_t|(\alpha + |Z|).
\end{aligned}$$

Using these bounds for the soft-thresholding function in (24), we have

$$\begin{aligned}
&\Big|\phi(\eta(B + \tau_* Z; \theta_*), B) - \phi(\eta(B + \tau_t Z; \theta_t), B)\Big| \\
&\leq L(1 + 4|B| + (\tau_* + \tau_t)|Z|)(\alpha + |Z|)|\tau_* - \tau_t|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\delta_t &= \mathbb{E}\Big|\phi(\eta(B + \tau_* Z; \theta_*), B) - \phi(\eta(B + \tau_t Z; \theta_t), B)\Big| \\
&\leq L|\tau_* - \tau_t|\mathbb{E}\Big[(1 + 4|B| + (\tau_* + \tau_t)|Z|)(\alpha + |Z|)\Big].
\end{aligned}$$

Using $\tau_* + \tau_t \leq 2\max\{\tau_0, \tau_*\}$, by the above $\delta_t \leq \kappa_2(\alpha, \tau_*)|\tau_* - \tau_t|$, where the constant $\kappa_2(\alpha, \tau_*) > 0$ does not depend on n or t and

$$\begin{aligned}
\kappa_2(\alpha, \tau_*) := L(\alpha + \mathbb{E}|Z|)(1 + 4\mathbb{E}|B|) \\
+ L2\max\{\tau_0, \tau_*\}(\alpha\mathbb{E}|Z| + 1).
\end{aligned} \quad (25)$$

The result follows from Lemma 3 with constants $K = \kappa_2 B_1$ and $\kappa = r_1$ where $r_1, B_1 > 0$ are defined in Lemma 3.

We now show $\widehat{e}_{t+1} \leq K e^{-\kappa(t+1)}$. By (11), we have $\lambda = \theta_*[1 - \frac{1}{\delta}\mathbb{E}[\eta'(B + \tau_* Z; \theta_*)]]$, and by the definition of the soft-threshold function in (4), we have that $\mathbb{E}[\eta'(x;\theta)] = P(|x| > \theta)$. Therefore,

$$\begin{aligned}
&\widehat{e}_{t+1} = \Big|\lambda - \theta_t\Big[1 - \frac{1}{\delta}\mathbb{E}[\eta'(B + \tau_t Z; \theta_t)]\Big]\Big| \\
&\leq \alpha|\tau_* - \tau_t| + \\
&\frac{\alpha\max\{\tau_0, \tau_*\}}{\delta}\Big|P(|B + \tau_* Z| > \theta_*) - P(|B + \tau_t Z| > \theta_t)\Big|
\end{aligned}$$

We have used that $|\theta_* - \theta_t| = \alpha|\tau_* - \tau_t|$ and that θ_* and θ_t are both upper bounded by $\alpha\max\{\tau_0, \tau_*\}$ since the sequence $\{\tau_t\}_{t\geq 0}$ is monotone by Lemma 1. Now the desired upper bound then follows by Lemma 3 and Lemma 6 in the supplementary materials.

The fact that $e_t \leq K e^{-\kappa t}$ follows using $\widehat{e}_t \leq K e^{-\kappa t}$ as proved above, $\theta_{t-1} \geq \alpha\min\{\tau_0, \tau_*\}$, and that $\tau_t^2 - 2E_{t,t-1} + \tau_{t-1}^2 \leq K e^{-\kappa t}$ by Lemma 3. The result $\nu_t^2 \leq K t e^{-\kappa t}$ follows immediately from $e_t \leq K e^{-\kappa t}$.

□

4 OPTIMIZATION LEMMA

The proof of Theorem 3 uses the following lemma, which gives conditions under which the result of Theorem 3 is true. Then to prove Theorem 3, we must prove that the conditions in Lemma 4 are met. We will state Lemma 4, leaving the proof to a longer version of this manuscript. Then we use the remainder of the paper to sketch a proof of the conditions.

Before stating the lemma, we introduce some additional notation. Consider a matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$ and a vector $\mathbf{v} \in \mathbb{R}^p$. Then for any subset $S \subset [p]$, we let $\mathbf{M}_S$ be the sub-matrix of $\mathbf{M}$ consisting of just the columns of $\mathbf{M}$ corresponding to the subset S and $\mathbf{v}_S$ be the vector of elements of $\mathbf{v}$ in subset S. We let the support of a vector be denoted $\text{supp}(\mathbf{v}) = \{i : v_i \neq 0\}$. We will consider the spectral properties of the $n \times p$ i.i.d. Gaussian matrix $\mathbf{X}$, labelling $\sigma_{min}^2(\mathbf{X})$ and $\sigma_{max}^2(\mathbf{X})$ the minimum and maximum singular values of the matrix $\mathbf{X}$. We label $\hat{\sigma}_{min}(\mathbf{X})$ to be the minimum, *non-zero* singular value of $\mathbf{X}$. Finally, the sub-differential of a convex function $f : \mathbb{R}^p \to \mathbb{R}$ at any point $\mathbf{x} \in \mathbb{R}^p$ is denoted by $\partial f(\mathbf{x})$. We refer frequently to the sub-differential of the ℓ_1 norm, $\|\mathbf{x}\|_1 = \sum_{i \in [n]} |x_i|$, as

$$\partial\|\mathbf{x}\|_1 = \{\mathbf{v} \in \mathbb{R}^p : |v_i| \leq 1\, \forall i; x_i \neq 0 \to v_i = \text{sign}(x_i)\}.$$

Lemma 4. *Suppose the following are true for generic constants* $K, \kappa \geq 0$.

Condition 1. *For some universal constant* $\mathsf{b}_1 = \max\{\hat{\mathsf{B}}, \tilde{\mathsf{B}}\} > 0$ *defined in Lemma 7 in the supplementary materials,*

$$P\Big(\frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2 \geq 4\mathsf{b}_1\Big) \leq K K_{t-1} e^{-\kappa\kappa_{t-1} n}.$$

Condition 2. *For constant* e_t *defined in Lemma 2 there exists a sub-gradient* $sg(C, \boldsymbol{\beta}^t) \in \partial C(\boldsymbol{\beta}^t)$ *with*

$$P\Big(\frac{1}{\sqrt{p}}\|sg(C, \boldsymbol{\beta}^t)\| \geq \epsilon + \sqrt{e_t}\Big) \leq K K_{t-1} e^{-\kappa\kappa_{t-1} n e_t \epsilon^2}.$$

Condition 3. *Let* $c_1, c_2,$ *and* c_3 *be universal con-*

stants with $0 < c_1, c_2, c_3 < 1$. *Define*

$$\boldsymbol{v}^t := (1/\lambda)[\mathbf{X}^*(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^t) + sg(C, \boldsymbol{\beta}^t)] \in \partial\|\boldsymbol{\beta}^t\|_1$$

and $S^t(c_1) := \{i \in [N] : |v_i^t| \geq 1 - c_1\}$ *where* $sg(C, \boldsymbol{\beta}^t)$ *is the sub-gradient from Condition 2 above. Then for any* $S' \subset [N]$ *with* $|S'| \leq c_2 N$, *it follows*

$$P(\sigma^2_{min}(\mathbf{X}_{S^t(c_1)\cup S'}) \leq c_3) \leq Ke^{-\kappa n}.$$

Condition 4. *For constants* $c_{min}, c_{max} > 0$, *the maximum and minimum non-zero singular values of* $\mathbf{X}$ *satisfy,*

$$P(\hat{\sigma}^2_{min}(\mathbf{X}) \leq c_{min} - \epsilon) \leq Ke^{-\kappa n\epsilon^2},$$

and

$$P(\sigma^2_{max}(\mathbf{X}) \geq c_{max} + \epsilon) \leq Ke^{-\kappa n\epsilon^2}.$$

Assume $t = O\Big(\frac{\log n}{\log\log n}\Big)$. *Then, for* $\epsilon \in (0, 1)$, *all* $t \geq t_{min}$ *(whose value isn't stated explicitly but doesn't depend on* ϵ *or* n*), and for large enough* n,

$$P\Big(\frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2 \geq \epsilon + \nu_t^2\Big) \leq KK_t e^{-\kappa\kappa_t n\epsilon^2}, \tag{26}$$

where e_t *is defined in* (20) *and*

$$\nu_t^2 = \frac{4e_t\sqrt{\mathsf{b}_1}}{c_{min}} + \Big(1 + \frac{tc_{max}}{c_3}\Big)\tilde{e}_t$$

with

$$\tilde{e}_t = \frac{32\sqrt{e_t\mathsf{b}_1}}{c_1^2 c_2}\Big(\frac{\sqrt{e_t\mathsf{b}_1}}{\lambda^2} + \frac{4}{c_{min}}\Big).$$

ν_t^2 *is also defined in Lemma 2 equation* (21), *though we state it again here to show the dependence on the other constants* $c_1 - c_3$ *in conditions (1) - (4) above.*

The result given by Lemma 4 is a refined result of [Bayati and Montanari, 2012, Lemma 3.1], the main difference being that our proof carefully tracks the relationship between t and n along with the rates at which critical values are concentrating. The details of the proof of Lemma 4 will be given in a longer version of this manuscript. The proof of Theorem 3 follows from Lemma 4 if one can show that Conditions 1-4 hold and we now sketch how one shows these results, leaving the full details to a longer manuscript.

The first condition follows from Lemma 7 in the supplementary material, which shows that the norms of the LASSO estimator, $\widehat{\boldsymbol{\beta}}$, and the AMP estimate at time t, $\boldsymbol{\beta}^t$, concentrate around constant values.

Condition 2 is proved by defining a specific subgradient that meets the requirement of the condition, and in particular, this subgradient depends on the differences $\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t$ and $\mathbf{z}^{t+1} - \mathbf{z}^t$, i.e. the differences in the output of the AMP algorithm (5) - (6) at subsequent iterations. The main technical piece is showing that the norms $\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2/p$ and $\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2/n$ concentrate on known values that approach 0 as t grows. This is mostly a consequence of the AMP analysis provided by [Rush and Venkataramanan, 2015].

The most difficult part of proving Theorem 3 is proving Condition 3. This requires proving some general statements about the minimum singular value of sub-matrices of $\mathbf{X}$ and also arguing that the sequence of sets considered by $S^t(c_1)$ does not change considerably when $t \geq t_{min}$. We note that t_{min} will be defined explicitly in the longer version of this document, and it depends on problem parameters related to the rate of algorithm convergence like B_1 and r_1 of Lemma 3, and we can ensure $t_{min} < T^*$ by allowing T^* large enough, though letting T^* grow degrades the rate of the concentration given in Theorem 1.

Condition 4 follows from Lemma 5 in the supplementary material, which gives concentration for the singular values of random matrices.

5 DISCUSSION

This work studies the asymptotic rate at which the loss between the LASSO estimator and the truth in a high-dimensional linear regression model approaches predicted values when $n/p \to \delta$ a constant $\delta \in (0, \infty)$, and the measurement matrix has i.i.d. Gaussian entries. An important tool in the proof is an approximate message passing (AMP) algorithm constructed to find the LASSO estimator and the theory relating to the AMP state evolution. In future work we hope to extend beyond the i.i.d. Gaussian measurement matrices, which will require a refined analysis of the AMP algorithms considered.

After providing a rigorous proof of the rate at which the AMP algorithm converges to the LASSO estimator, we use this convergence to pass along the AMP algorithm state evolution guarantees to provide asymptotic predictions about the LASSO loss. A limitation of this work is that the current finite sample AMP state evolution analysis requires the number of iterations t for which the algorithm is run, to be such that $t = O\Big(\frac{\log n}{\log\log n}\Big)$. Hence, when considering the rates at which the LASSO loss concentrates to the predicted values, this rate holds.

Acknowledgements

The author would like to thank Yanting Ma and Yumming Zhang for their helpful discussions relating to this work. This work was supported in part by NSF CCF #1849883.

References

[Bayati and Montanari, 2011] Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory*, pages 764–785.

[Bayati and Montanari, 2012] Bayati, M. and Montanari, A. (2012). The LASSO Risk for Gaussian Matrices. *IEEE Trans. Inf. Theory*, 58(4):1997–2017.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford.

[Bu et al., 2019] Bu, Z., Klusowski, J., Rush, C., and Su, W. (2019). Algorithmic analysis and statistical estimation of slope via approximate message passing. *arXiv preprint arXiv:1907.07502*.

[Chen and Donoho, 1995] Chen, S. and Donoho, D. (1995). Examples of basis pursuit. In *Wavelet Applications in Signal and Image Processing III*, volume 2569, pages 564–574. International Society for Optics and Photonics.

[Donoho et al., 2009a] Donoho, D., Maleki, A., and Montanari, A. (2009a). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

[Donoho et al., 2009b] Donoho, D., Maleki, A., and Montanari, A. (2009b). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

[Donoho and Montanari, 2016] Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via Approximate Message Passing. *Probability Theory and Related Fields*, 166(3-4):935–969.

[Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[Krzakala et al., 2012] Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, (8).

[Montanari, 2012] Montanari, A. (2012). Graphical models concepts in compressed sensing. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing*, pages 394–438. Cambridge University Press.

[Rangan, 2011] Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 2168–2172.

[Rangan et al., 2019] Rangan, S., Schniter, P., and Fletcher, A. (2019). Vector approximate message passing. *IEEE Transactions on Information Theory*.

[Rush and Venkataramanan, 2015] Rush, C. and Venkataramanan, R. (2015). Finite sample analysis of approximate message passing. *Proc. IEEE Int. Symp. Inf. Theory*. Full version: `https://arxiv.org/abs/1606.01800`.

[Su et al., 2017] Su, W., Bogdan, M., Candes, E., et al. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, 45(5):2133–2150.

[Sur and Candès, 2019] Sur, P. and Candès, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

[Thrampoulidis et al., 2015] Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Zheng et al., 2017] Zheng, L., Maleki, A., Weng, H., Wang, X., and Long, T. (2017). Does ℓ_p-minimization outperform ℓ_1-minimization? *IEEE Transactions on Information Theory*, 63(11):6896–6935.